```
- text in red
+ text in green
! text in orange
# text in gray
@@ text in purple (and bold) @@
```

Marked text

Marked text

Some green text

We can use NLP to create systems like speech recognition, document summarization, machine translation, spam detection, named entity recognition, question answering, autocomplete, predictive typing and so on.

<⟶**>** What is Corpus ?

A Corpus is defined as a collection of text documents for example a data set containing news is a corpus or the tweets containing Twitter data is a corpus.

Corpus ⟶ Documents ⟶ Paragraphs ⟶ Sentences ⟶ Tokens

Tokens can be words, phrases, or Engrams, and Engrams are defined as the group of n words together.

For example, consider this given sentence-
“I love my phone.”
uni-grams(n=1) are: I, love, my, phone
Di-grams(n=2) are: I love, love my, my phone
And tri-grams(n=3) are: I love my, love my phone

<⟶**>** What is Tokenization?

Tokenization is a process of splitting a text object into smaller units called tokens.
The most commonly used tokenization process is **White-space Tokenization**.

<→**>** What is White Space Tokenization?

Also known as **unigram tokenization**. Entire text is split into words by splitting them from white spaces.

For example, in a sentence- “I went to New-York to play football.”
This will be splitted into following tokens: “I”, “went”, “to”, “New-York”, “to”, “play”, “football.”

<→**>** What is Regular Expression Tokenization ?

A regular expression pattern is used to get the tokens.. We can split the text by passing a splitting pattern.

```
 Sentence= “Football, Cricket; Golf Tennis"
 re.split(r’[;,\s]’, Sentence)
```

Tokens= “Football”, ”Cricket”, “Golf”, “Tennis”

Tokenization can be performed at the sentence level or at the world level or even at the character level

<→**>** What is Normalization?

A Morpheme is defined as the base form of a word.A token is generally made up of two components, Morphemes, which are the base form of the word, and Inflectional forms, which are essentially the suffixes and prefixes added to morphemes.
Eg: Antinationalist: Anti + national + ist ⟺ prefix + morpheme + suffix

Normalization is the process of converting a token into its base form. Helps in removing redundant information
Few normalization techniques are **Stemming and Lemmatization.**

<→**>** What is Stemming?

Rule-based process to remove inflection part from a token. **Output is a stem**. They are easier to implement and usually run faster. Stemmer operates without knowledge of the context. Crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time.

Eg: "laughing","laughed","laughs","laugh" >>> "laugh"

Not Good as can produce words that are not in the dictionary
Eg: “His teams are not winning”
After stemming the tokens that we will get are- “hi”, “team”, “are”, “not”, “winn”
Notice that the keyword “winn” is not a regular word and “hi” changed the context of the entire sentence.

<—> What is Lemmatization?
Lemmatization is a systematic step-by-step process for removing inflection forms of a word. **Output is a lemma**. Uses vocabulary and morphological analysis of words.
Eg: Running, Ran, Run >> Run

<—> Parts of Speech Tags in Natural Language Processing

PoS tags is the properties of words that define their main context, their function, and the usage in a sentence.
Some of the commonly used parts of speech tags are- Nouns, which define any object or entity; Verbs, which define some action; and Adjectives or Adverbs, which act as the modifiers, quantifiers, or intensifiers in any sentence.
They are used in a variety of tasks such as text cleaning, feature engineering tasks, and word sense disambiguation.

<—> Grammar in NLP

- Constituency Grammar
- Dependency Grammar

What is Constituency Grammar?
Defines the structural pieces of a sentence, phrase, or clause driven by driven by their part of speech tags, noun or verb phrase identification.
Subject + Context + Object
“The dogs are barking in the park.”
“They are eating happily.”
“The cats are running since morning.”

What is Dependency Grammar?

Words of a sentence are dependent upon other words of the sentence.

I Didnt understand this

Used for Named Entity Recognition, Question Answering System, Coreference Resolution, Text summarization and Text classification

What are Stop Words?

Stop words are filtered out before or after processing of text as they can add lot of noise. Eg: "and", “the”, “a”